Prediction of Total Knee Replacement Using Radiographs and Clinical Risk Factors

Jack Jones, Gary Ng, Sanjay Subramanian
Center for Data Science
New York University
New York, NY 10011

Abstract

Accurate assessment of progression of knee osteoarthritis (OA) is essential for improving clinical outcomes. Current methods do not provide enough information to make accurate and robust outcome predictions. Here, we develop traditional and deep learning (DL) models to assess risk of OA progression to total knee replacement (TKR) by analyzing knee radiographs and clinical risk factor data. We attempt to extract the most important features from these models as part of a combined model to utilize all available data, both structured and unstructured. Our results suggest that a combined model using logits of the predicted probabilities of the individual traditional and DL models as inputs outperforms the individual models and report an AuROC of 0.906 and AuPRC of 0.469.

1 Introduction

Osteoarthritis (OA) is one of the most prevalent chronic diseases in the US and worldwide. The knee is the joint most commonly affected with OA. Knee OA is characterized by joint pain, stiffness, and decreased range of motion. In the US, over 14 million patients are diagnosed with knee OA each year, with the prevalence rate progressively increasing with the increasing age of the American population [1]. Knee OA is also a leading cause of chronic disability in the US and a major socioeconomic burden on American society. Total knee replacement (TKR) is the only effective treatment option for patients with advanced knee OA. More than half of patients diagnosed with knee OA will undergo primary TKR during their lifetime, with over 600,000 surgical procedures performed each year in the US alone [2].

There is an important clinical need to identify patients with knee OA at risk for progression to TKR. Identifying patients with early knee OA at high risk for TKR could provide a window of opportunity for intervention during the early stages of the disease process, when conservative treatment options are most likely to be successful. Identifying these patients could also improve the decision-making process for surgical treatment, which must be performed before comorbid medical conditions make the surgical procedure unsafe [3]. Current traditional risk assessment models rely on a priori extracted clinical and radiographic risk factor data, including age, race, sex, body mass index, (BMI), history of knee injury, and Kellgren-Lawrence (KL) grade of radiographic knee OA. However, these traditional risk assessment models have shown only moderate success in predicting progression to TKR, making it necessary to utilize more powerful automated tools for this task.

Deep learning (DL) is well suited for creating OA risk assessment models as it provides a fully automated method to extract prognostic information from representative subsets of features on baseline imaging studies. In this study, we aim to combine tabular clinical risk factor data with radiograph data to obtain a more robust prediction of TKR on a subset of subjects in the Osteoarthritis Initiative (OAI) longitudinal study, consisting of 8,932 knees from 4,511 unique subjects with a prediction label indicating whether each knee had undergone TKR over a nine-year observational

Preprint. Under review.

period. We hypothesize that joint ensemble models leveraging all available data will have higher diagnostic performance for predicting TKR compared to current traditional and DL models.

2 Related Work

Predictive models for TKR have a limited history, with much of the existing methodology analyzing baseline clinical and radiographic risk factors as opposed to images. In addition, measures of knee OA progression other than TKR, including pain progression and radiographic progression, have been widely used as prediction labels, which can lead to inconsistencies and a lack of model reproducibility given their subjectivity. Many of the clinical and radiographic risk factors used in current models for predicting knee OA progression vary from institution to institution and are limited by this subjectivity and lack of consistency. For example, KL grade of radiographic knee OA, which assigns a numerical score between 0 to 4 to assess overall disease severity on X-rays, is one of the strongest risk factors for knee OA progression. However, there is only moderate inter-reader agreement between experienced musculoskeletal radiologists for assigning a KL grade for knee OA.

Previous studies have described risk assessment models for predicting knee OA progression to TKR. Most studies have used random forest regression and t-tests to identify clinical and radiographic risk factors most strongly associated with TKR or have created traditional OA risk assessment models combining these clinical and radiographic risk factors using linear regression or machine learning approaches. A few previous studies have described DL models for predicting TKR analyzing baseline imaging studies and have shown that these DL models outperform traditional models analyzing baseline clinical and radiographic risk factors. Tolpadi et.al. reported that DL models created from 3,400 knees in the OAI analyzing baseline X-rays alone, baseline MRI alone, a combination of baseline risk factors and DL analysis of X-rays, and a combination of baseline risk factors and DL analysis of MRI had high diagnostic performance for predicting TKR over an eight-year observational period, with AuROCs of 0.848, 0.886, 0.890, and 0.834, respectively [4]. Leung et. al. performed two nested case-control studies on a subset of 718 knees in the OAI and reported that DL models analyzing baseline X-rays and MRI alone had AUCs of 0.870 and 0.866 respectively for predicting TKR over an eight-year observational period [5].

A large portion of these models rely on AuROC as a final evaluation metric. However, most available data present with major class imbalance, as the proportion of patients who undergo TKR is generally tied to patients with advanced OA and is thus significantly lower than the proportion of patients who do not undergo TKR. In this case, the positive class is important to identify accurately in order to minimize false negatives, a distinction which traditional AuROC may not be able to robustly assess.

3 Problem Definition

The goal of our study is to develop reproducible and robust risk assessment models for predicting knee OA progression to TKR. These models could identify high risk patient populations that could be treated with conservative interventions during the earliest stages of knee OA, which could ultimately reduce healthcare costs and improve clinical outcomes. These models would also be useful in the decision-making process for TKR, an elective procedure, as patients with advanced knee OA contemplate the most appropriate timing for surgical treatment.

3.1 Task

Our modeling approach is divided into three segments, composed of the traditional model, DL model, and combined ensemble model. The traditional model takes the structured clinical and radiographic risk factor data as input, while the DL model takes radiograph images as input. The ensemble model concatenates the outputs of these individual models in order to leverage all available data, both structured and unstructured.

3.2 Approach

A single validation set approach is used to train the individual models, followed by a four-fold cross validation strategy applied to the combined model. Logistic regression is chosen as an initial baseline to be trained on a set of core features that are known and widely available predictors or

indicators of OA progression. Logistic regression offers an interpretable standard for benchmarking the following models. For our clinical data, we use a gradient boosted approach, specifically CatBoost, which is well-suited for high-dimensional tabular data and allows us to compute various types of feature importance in order to trim excess features and improve interpretability and runtime. Our proposed DL model adapts a transfer learning approach based on the well-documented Resnet34 architecture, with the number of output features in the final fully connected layer changed from 1,000 to 2. Pre-trained models such as Resnet offer more complex feature extraction layers and leverage training on millions of images while including skip connections to allow for stacking of layers without any degradation in performance. In order to utilize the information output from both the structured CatBoost and unstructured CNN models, the ensemble framework combines predictions into a single model through various methods, including a concatenation of the logits of the outputs combined with various permutations of the other existing features from the individual models. Given the imbalanced nature of the data, the metric of choice in this case is the area under the precision-recall Curve (AuPRC). The traditionally reported metric AuROC can be misleading in imbalanced cases, outputting a high score while misclassifying much of the minority class. In this case, it is imperative to avoid false negatives so that all potential future cases of TKR may be identified and mitigated. Unlike AuROC, where the baseline in every case is 0.5, baseline AuPRC is equal to the fraction of positive cases, and this metric is thus generally lower in value than AuROC.

4 Experimental Evaluation

4.1 Traditional Model

4.1.1 Data

We utilize three clinical datasets from the Osteoarthritis Initiative (OAI):

- *Enrollees_SAS* is a patient-level, visit-independent dataset that contains demographic and other grouping information about patients that are not expected to change from year to year
- *AllClinical00_SAS* combines a number of patient-level datasets containing self-reported medical history, nutritional intake, physical exam measurements, and other information at the time of baseline visit
- *kXR_SQ_BU00_SAS* is a knee-level dataset containing readings of patients' X-rays, including the Kellgren-Lawrence (KL) grade, a classification of osteoarthritis severity known to be a predictor of TKR

In addition to the knee-level *kXR_SQ_BU00_SAS* dataset, some features in the patient-level *AllClinical00_SAS* dataset also reveal information about specific knees. For instance, $\mathrm{V00LKRFXPN}$ indicates pain/tenderness was present on the patient's left knee during examination, whereas $\mathrm{V00RKRFXPN}$ indicates similar information about the patient's right knee. We identified and matched up 192 pairs of these left/right knee-specific "patient-level" questions programmatically, and joined them back to our knee-level observations as two sets of features: one set pertaining to the knee in question, and another describing the other knee of the same patient. We posit that since the decision to undergo knee replacement on either knee is made by the same individual, information about one knee might be predictive of the decision on the other.

Truly patient-level features such as demographic information and nutritional intake are joined to our knee-level observations by patient id, so that both knees of the same patient share the same feature values. The resulting dataset contains nearly 2K features. We elect to drop features with more than 10% null values, which leave us with 685 eligible features to build our traditional model.

4.1.2 Methodology

For each model below, we set aside 25% of the non-holdout data as our validation set, to help us make various decisions about the model, including feature selection, early-stopping, hyperparameter-tuning, and other design choices.

Baseline Model (1): Logistic Regression with Core Features As a baseline model, we fit a logistic regression using a set of core features, namely $\mathrm{KLGrade}$, BMI, Age, Sex, Race, $\mathrm{KneeInjury}$, and

KneeAlignment that are known predictors of osteoarthritis progression commonly used in other research studies. Additional steps are performed to help the linear model learn effectively: (i) categorical features like Race and Sex are dummified, (ii) squared terms of numerical features Age and KneeAlignment are added in case their effects on the response variable are non-linear, (iii) attempt to drop statistically insignificant features based on p-values. These steps are performed iteratively and retained if and only if it improves validation AuPRC. The regression output of the resulting logistic regression is shown in Figure 4 of the Appendix. Notably, (i) of the dummy race terms, only RaceIsBlack is statistically significant and dropping all others, effectively reducing the categorical race feature into a binary one, significantly improves validation AuPRC; (ii) adding squared term of Age helped, but not of KneeAlignment; (iii) while KneeInjury and KneeAlignment features are deemed statistically insignificant by p-values, removing them worsens validation AuPRC and they are thus retained in the final model.

Baseline Model (2): CatBoost with Core Features Logistic regression is not well-suited to learn from our full dataset of hundreds of features because it (i) does not capture interaction between features, (ii) does not accept null values, of which many features do have, and (iii) assumes linearity between the features and the target variable. Granted, it is possible to overcome these limitations with careful treatment of the features, as we have done in our baseline model with the core features, but this approach is hardly scalable given the number of features we have. Instead, our algorithm of choice is gradient boosting, which capture non-linear and interaction effects well, and arguably remains state-of-the-art for machine learning with tabular data. Specifically, we choose to fit our models using the CatBoost library, an increasingly popular alternative to XGBoost, primarily due to its ease of use, which in turns allows us to iterate faster. For instance, it does not require categorical features to be pre-processed, nor does it require extensive tuning to achieve good performance because its default hyperparameters tend to work very well out-of-the-box. It also provides convenient methods to plot learning curves, extract SHAP values, and compute various types of feature importances. We first fit CatBoost using the same core features in our logistic regression as our second baseline model.

Final Model: CatBoost with Expanded Features Finally we expand our feature space and build a more complex CatBoost model using the full set of eligible features. However, it is highly likely that only a subset of the 685 features are informative, and that we can build a more robust model with better out-of-sample predictive performance by dropping redundant features and reducing noise in our model. To this end, we rely on the `get_feature_importance(type="LossFunctionChange")` method of CatBoost, which approximates the change in loss function when a feature is included vs. excluded. We begin by fitting a full model with all features, and then iteratively subsequent models by refitting with only features that yielded positive importances, until there remains no more features with zero or negative importances to drop. This yields a series of increasingly parsimonious CatBoost models, and we pick the model with a subset of features that had the lowest validation loss as our final model. In total, 44 features are retained in our final CatBoost model, and they are listed along with their feature importances in Table 6 of the Appendix.

4.1.3 Results

The final CatBoost model and its baselines are evaluated on the hold-out test set, with results shown below in Table 1. The final CatBoost model significantly outperforms both baseline models in AuPRC as well as AuROC and LogLoss. Note however that the CatBoost baseline model does not perform meaningfully better on AuPRC, and in fact does worse on AuROC and LogLoss, than the logistic regression baseline model when given access to only the same set of core features, but this is not really surprising because a low-dimensional feature space tends not to demand or benefit from the flexibility that more complex models offer. Perhaps more importantly, the huge improvement between the baseline CatBoost model and the final model underscores that there are other clinical risk factors beyond the core factors that signal whether a patient will undergo total knee replacement.

Table 2 below shows the top 10 features in the final CatBoost model ranked by their `LossFunctionChange` feature importances (the full list may be found in Table 6 of the Appendix). Also tabulated alongside these feature importances in Table 2 are the `ShapImportance`, which we define to be each feature's mean absolute SHAP values in the validation set. SHAP values are computed at the observation/feature level, describing a given feature's contribution in dragging the prediction value away from the mean prediction (positive SHAP means the feature increased the predicted probability of the observation).

Table 1: Test set accuracy metrics of traditional models using clinical data

Model	LogLoss	AuROC	AuPRC
Baseline Model (1): Logistic Regression w/ Core Features	0.179	0.850	0.266
Baseline Model (2): CatBoost w/ Core Features	0.186	0.828	0.267
Final Traditional Model: CatBoost w/ Expanded Features	0.170	0.875	0.373

In Figure 1 the beeswarm plot of the SHAP values, for instance, shows that the older the patients (higher Age values), the more likely they are to undergo TKR (positive SHAP values). Aggregating the absolute values of SHAP values across the dataset gives us an alternative reading of feature importances (`ShapImportance`). Specifically, it tells us what features had the most impact on the prediction values output by our CatBoost model. Note that while `LossFuncChgImportance` and `ShapImportance` are highly correlated, they do not necessarily yield the same top features since they are measuring slightly different things. For example, Age has the highest `ShapImportance` but only ranks ninth in `LossFuncChgImportance`.

Between the two feature importances, however, we regard `LossFuncChgImportance` as the more useful insofar as telling us whether including the feature helps improve our model predictions. By this measure, KLGrade is by far the most important feature, followed by other baseline X-ray readings such as BUkXRReading, OsteophytesJSN, and CompositeOAGrade. Another group of features signaling the degree of pain and discomfort experienced by the patient, including FlexionPain, AwarenessKneeProblems, UsePainMedsL12M and KneeCatchHangUpL7D also rank high. This makes sense because TKR is to some degree an elective procedure, so some patients might choose to undergo it only if they find the pain unbearable or feel their lives severely impeded by it. It is also worth noting that RaceIsBlack ranks as an important feature in both our final CatBoost and baseline logistic regression models, with patients who identify as Black or American-American being less likely to undergo TKR. To the extent that race is correlated with income, this could indicate medical access and affordability issues that should be addressed and tackled in other research and initiatives.

Table 2: Feature importances and SHAP values of select features in final CatBoost Model

Feature	LossFuncChgImportance	ShapImportance
KLGrade	0.0137	0.3311
BUkXRReading	0.0035	0.0863
Osteophytes&JSN	0.0033	0.1886
CompositeOAGrade	0.0023	0.1221
FlexionPain	0.0022	0.0789
RaceIsBlack	0.0020	0.1012
AwarenessKneeProblems	0.0019	0.2831
UsedPainMedsL12M	0.0017	0.1829
Age	0.0017	0.3858
KneeCatchHangUpL7D	0.0015	0.0678

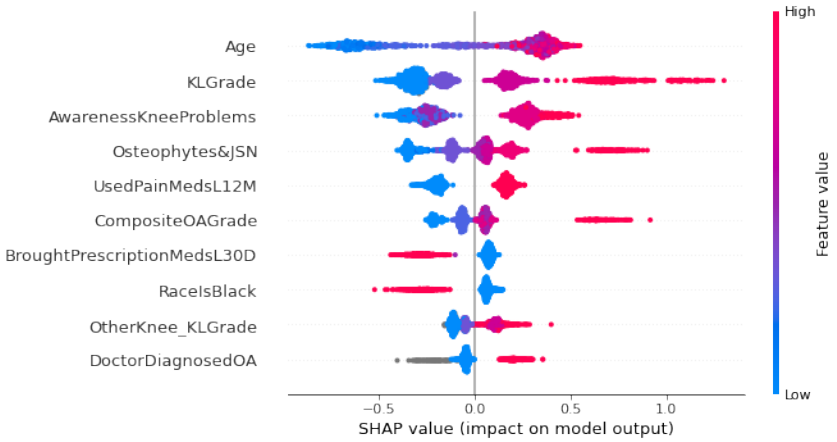

Figure 1: Beeswarm plot of SHAP values of select features in final CatBoost model

4.2 X-Ray Model

4.2.1 Data

We utilize 8,932 knee X-Ray images from 4,511 unique patients in the OAI longitudinal study. The X-rays are cropped to zoom in on the knee and stored in HDF5 files. All preprocessing is completed by prior study [6]. We further downsample the images to 256 x 256 and convert them to JPEG files for performance reasons prior to training.

4.2.2 Methodology

Similar to the Traditional Model approach, we set aside 25% of the non-holdout data as our validation set to experiment with during hyperparameter-tuning. Our final X-ray model is evaluated on a hold-out test set containing 2,233 knees.

X-Ray Model: Convolutional Neural Network We utilize the image classification model ResNet34, a 34 layer convolutional neural network (CNN), with weights that are pre-trained on ImageNet. The only modification made to the architecture is in the fully connected layer whose number of output features is changed from 1,000 to 2. We then train all the weights of the network on the X-ray data to minimize cross entropy loss where estimated probabilities are computed using softmax applied to the output of the network. Prior to each epoch of training, randomized data augmentations are applied to prevent overfitting. Images are randomly rotated, cropped to 224 x 224, and flipped horizontally. Color jitter is applied followed by a Gaussian blur. An example of a transformed X-ray image is shown in Figure 5 of the Appendix. Hyperparameter tuning is performed using the aforementioned validation set approach. The tuned parameters include learning rate, number of epochs, batch size, L2 regularization coefficient, learning rate decay, and choice of optimization procedure.

4.2.3 Results

The final X-ray model is evaluated on the hold-out test set with the results shown below in Table 3. This model outperforms the baseline logistic regression with core features in AuPRC and log loss, but not in AuROC. It does not outperform the final traditional model in any of the three evaluation metrics. There is an inherent bias with the validation set approach, as the model can overfit the validation set during the hyperparameter tuning process. A k-fold cross validation approach with early stopping could perform better on the global hold-out test-set. Additionally, the model with full resolution X-ray images (stored as HDF5 files) as inputs suffered from extensive runtimes, necessitating a conversion of the images to the JPEG file format.

Table 3: Test set accuracy metrics of X-ray model using X-ray data

Model	LogLoss	AuROC	AuPRC
X-ray Model: Covolutional Neural Network	0.211	0.795	0.283

4.3 Combined Model

While both the Traditional CatBoost model and the Xray CNN model yield strong out-of-sample performance, their predictions are only weakly correlated, with a correlation coefficient of 0.185. This suggests that the two models extract different information from the two datasets, clinical risk factors and X-ray images, and that combining the two might yield even better results.

4.3.1 Data

The primary inputs to our combined models are the predicted probabilities output by the aforementioned Traditional CatBoost model ($\mathrm{ClinicalPred}$) and the Xray CNN model ($\mathrm{XrayPred}$). In order to prevent data leakage, we derive out-of-sample predictions $\mathrm{ClinicalPred}$ and $\mathrm{XrayPred}$ for the training set by running a 4-fold cross validation on their respective models.

In addition to using the outputs from the Traditional CatBoost model and Xray CNN model as input features into our combined model, we also experiment with incorporating the input features from the Traditional CatBoost model, i.e. the clinical risk factors, directly in our combined model.

4.3.2 Methodology

We fit a number of combined models on various subsets of the aforementioned input features (and their derivatives) using logistic regression and CatBoost. We evaluate these models using a 4-fold cross validation on the full training set, and pick the model with the highest AuPRC as our best combined model:

LogitWithPreds A logistic regression using only $\mathrm{ClinicalPred}$ and $\mathrm{XrayPred}$ as input features. This is probably the most obvious way to combine the prediction output from our two models.

LogitWithPredLogits A variation of LogitWithPreds, this is a logistic regression using logits of the predicted probabilities, i.e. $\mathrm{logit}(\mathrm{ClinicalPred})$ and $\mathrm{logit}(\mathrm{XrayPred})$ as input features. Mathematically, the logit function is the inverse of the sigmoid function ("S-curve"), projecting the predicted probabilities from [0,1] onto scores with an unbounded range. Fitting our logistic regression on these scores instead of probabilities allows either model to pull the ensemble prediction farther in its direction if it is extremely confident.

CatBoostPredsOnly A CatBoost classifier using only $\mathrm{ClinicalPred}$ and $\mathrm{XrayPred}$ as input features. This is another variation of LogitWithPreds, except swapping out logistic regression with CatBoost as the classifier algorithm. There might be interaction and non-linear effects that CatBoost might better capture.

CatBoostFullWithXrayPreds A CatBoost classifier using $\mathrm{XrayPred}$ and the 44 risk factors selected for the Traditional CatBoost model. The idea here is to extend our Traditional CatBoost model by simply adding $\mathrm{XrayPred}$ another feature.

CatBoostFullWithBothPreds A CatBoost classifier using $\mathrm{XrayPred}$, $\mathrm{ClinicalPred}$, and the 44 risk factors selected for the Traditional CatBoost model. This is a variation of CatBoostFullWithXrayPreds, where we toss in the output of the Traditional CatBoost model as well. There is considerable redundancy in this model so it is unlikely to perform well, but it is a starting point for the next model.

CatBoostSelectedWithPreds This is a "pruned version" of CatBoostFullWithBothPreds, where we use only a subset of its features. We obtain this model using a similar iterative process as the

feature selection for the Traditional CatBoost model, i.e. sequentially fitting a series of increasingly-parsimonious models by dropping features with negative or zero importances, and then picking the one with the lowest validation loss (on a single fold of training data).

4.3.3 Results

Evaluated using cross-validation on the train set, all of the combined models described above perform better than either of the standalone models across all three evaluation metrics. The model with the highest validation AuPRC is LogitWithPredLogits, so we choose this as our final combined model.

Table 4: Train set CV accuracy metrics of combined models

	LogLoss	AuROC	AuPRC
LogitWithPreds	0.155	0.894	0.500
LogitWithPredLogits	0.141	0.901	0.571
CatBoostPredsOnly	0.144	0.898	0.545
CatBoostFullWithXrayPreds	0.141	0.909	0.559
CatBoostFullWithBothPreds	0.141	0.907	0.555
CatBoostSelectedWithPreds	0.142	0.900	0.562
Baseline (1): TraditionalOnly	0.161	0.878	0.427
Baseline (2): XrayOnly	0.197	0.787	0.249

The regression summary in Figure 2 below shows $\mathrm{ClinicalPredLogit}$ with a higher coefficient than $\mathrm{XrayPredLogit}$, corroborating the fact that the standalone Traditional CatBoost model performed better than the standalone X-ray CNN model.

Logit Regression Results

Dep. Variable:	KR_LABEL	**No. Observations:**	6695
Model:	Logit	**Df Residuals:**	6692
Method:	MLE	**Df Model:**	2
Date:	Sun, 05 Dec 2021	**Pseudo R-squ.:**	0.4028
Time:	18:07:41	**Log-Likelihood:**	-916.21
converged:	True	**LL-Null:**	-1534.1
Covariance Type:	nonrobust	**LLR p-value:**	4.549e-269

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.4104	0.174	13.842	0.000	2.069	2.752
ClinicalPredLogit	1.1427	0.049	23.101	0.000	1.046	1.240
XrayPredLogit	0.7612	0.046	16.459	0.000	0.671	0.852

Figure 2: Regression output for Combined Model LogitWithPredLogits

Finally, we evaluate our chosen combined model on the test set and see that it obtains an AuPRC of 0.469 and an AuROC of 0.906 (95% confidence interval of 0.884-0.927), and logloss of 0.158, outperforming both the Traditional CatBoost model and X-ray CNN model by a considerable margin. Using the DeLong test to compare areas under correlated ROC curves, we determine that difference between the combined model's and Traditional CatBoost model's AUC is statistically significant (p-value<0.001). In addition, the Youden index was used to determine the optimal model sensitivity and specificity for each model, as displayed in Table 5 and annotated with open circles on the ROC Curves in Figure 3. The combined model achieves a good balance between sensitivity (0.846) and specificity (0.812), both of which are also higher than those of the standalone models.

Table 5: Test accuracy metrics of final models

	LogLoss	AuPRC	AuROC (95% CI)	Sensitivity	Specificity
Combined Model	0.158	0.4691	0.906 (0.884 - 0.927)	0.846	0.812
Clinical CatBoost Model	0.170	0.3732	0.875 (0.849 - 0.901)	0.831	0.763
X-ray CNN Model	0.211	0.2825	0.795 (0.754 - 0.836)	0.757	0.727

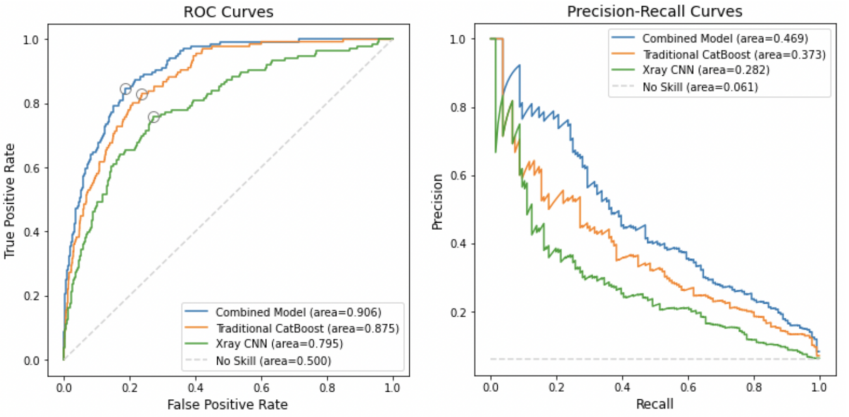

Figure 3: ROC and Precision-Recall Curves

5 Discussion

In this study, we find that all six variations of our combined model have a significantly higher diagnostic performance than our standalone traditional and DL models when assessed using AuPRC, AuROC, and log loss. The model using the combined logits of the predicted probabilities from each of the standalone models yields the best out-of-sample performance of these ensembles, with an AuPRC of 0.469, AuROC of 0.906, and log loss of 0.158. In addition, this model outperforms the standalone models in sensitivity and specificity, with values of 0.846 and 0.812 respectively. This model also outperforms results reported in current literature, which present a maximum AuROC of 0.870 when predicting OA progression to TKR using clinical risk factor data and radiographs via a DL framework.

The novelty in this method for predicting OA progression to TKR lies in the combined logit framework, which outputs robust results without much complexity involved in its structure or input selection. Using the logits of the probability outputs from the standalone models rather than the probabilities themselves allows the model to sway the combined output in either direction more confidently based on the magnitude of these probabilities. These results show extreme promise for deployment in a clinical setting, as there is a growing need for clinical interventions that prevent or delay the need for TKR. Having access to a powerful indicator of potential surgery occurrence in patients based on baseline knee radiographs would allow for personalized, specific, and early interventions that would lead to better outcomes and reduced costs in the long term.

Some limitations also exist in the implementation of this model, with room for further refinement. Given the three-pronged nature of this study, improvements to either of the standalone models could potentially result in immediate enrichment of the combined model. In the traditional CatBoost model, a number of hyperparameters such as learning rate and depth could be trained more extensively rather than relying on default values, and a principal component analysis (PCA) could be performed to reduce dimensionality of the feature space and filter out heavily correlated variables. This approach could improve performance of the model at the expense of interpretability. The DL model, which was outperformed by the CatBoost model, suffered from extensive runtimes when trained on full resolution radiographs. While downsizing the images and converting them to the JPEG format expedited this process, future iterations could seek to obtain the full resolution of baseline images without compromising and bottlenecking runtime. In addition, more complex pre-trained architectures such as Resnet50 or DenseNet could boost performance by adding additional layers of complexity.

The training process for these models was performed using a single validation set. This could be made more robust through a k-fold cross validation framework, which could yield better results on the eventual global hold-out test-set.

TKR is an elective procedure, so clinical risk factors such as affordability and pain tolerance may hold additional weight over radiographs when compared to studies predicting structural OA progression as an outcome. Modulating the prediction label to forecast joint space loss rather than TKR could provide additional information about progression and reveal more underlying value in the images.

In conclusion, our study has demonstrated the feasibility of using joint traditional and DL risk assessment models for predicting the progression of knee OA to TKR over a nine-year follow-up period using baseline knee radiographs and clinical and radiographic risk factor data. Joint models have a significantly higher performance for predicting OA progression to TKR when compared to individual traditional and DL methods. While further hyperparameter tuning and validation stages are needed, this study provides a promising framework for joint techniques that are able to leverage all available data from baseline studies.

6 Lessons Learned

As is often the case with machine learning and especially deep learning projects, data processing and formatting and model runtime were bottlenecks to completion. In this project, data appropriation from prior studies and repositories involved reorganizing thousands of images based on patient ID and knee sidedness and stratifying these into equivalent training, validation, and testing splits for consistency across all models. In addition, the size of full resolution radiograph images significantly delayed DL model training times, making hyperparameter tuning and experimentation difficult. Alleviating this by downsizing and compressing images expedited this training process by orders of magnitude while still providing robust results. Application of additional data manipulation techniques for augmentation (rotations, transformations, addition of jitter or noise, etc.) is a technique widely utilized in DL models dealing with images as it adds stochasticity into networks and enhances learning ability. Employing the Resnet architecture also allowed for a more interpretable DL model, as this is a network that has been heavily studied and standardized across many fields. Being able to make minor modifications to fit the needs of our study simplified the methodology without sacrificing output or efficiency. External resources utilized for processing power are generally required when running these types of models when local units do not suffice. Google Colab GPU units were used to run models in batches and optimize training time. Each of these adjustments are important to understand in the future, especially in projects dealing with multi-modal data.

7 Acknowledgments

We would like to thank our mentors Dr. Cem Deniz, Ph.D. and Dr. Richard Kijowski, M.D. of the Department of Radiology at the NYU Grossman School of Medicine, and Elena Sizikova of the NYU Center for Data Science.

8 Student Contributions

JJ: data exploration, data processing, X-ray model tuning, final report
GN: data exploration, data processing, traditional model, tuning, combined model implementation, final report
SS: data exploration, literature review, poster content, final report

References

[1] Maradit Kremers, H. et al. "Prevalence of Total Hip and Knee Replacement in the United States." *The Journal of bone and joint surgery*. American volume vol. 97,17 (2015): 1386-97. doi:10.2106/JBJS.N.01141.

[2] Inacio, M.C.S. et al. "Projected increase in total knee arthroplasty in the United States - an alternative projection model." *Osteoarthritis and cartilage* vol. 25,11 (2017): 1797-1803. doi:10.1016/j.joca.2017.07.022.

[3] Guan, B. et al. “Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-MONTH follow-up period.” *Osteoarthritis and cartilage* vol. 28,4 (2020): 428-437. doi:10.1016/j.joca.2020.01.010.

[4] Tolpadi, A.A. et al. “Deep Learning Predicts Total Knee Replacement from Magnetic Resonance Images.” *Sci Rep* vol. 10, 6371 (2020). https://doi.org/10.1038/s41598-020-63395-9.

[5] Leung, K. et al. “Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative.” *Radiology* vol. 296,3 (2020): 584-593. doi:10.1148/radiol.2020192091.

[6] Zhang, B. et al. (2020). “Attention-based CNN for KL Grade Classification: Data from the Osteoarthritis Initiative.” *ISBI 2020 - 2020 IEEE International Symposium on Biomedical Imaging* (pp. 731-735). (Proceedings - International Symposium on Biomedical Imaging; Vol. 2020-April). IEEE Computer Society. https://doi.org/10.1109/ISBI45749.2020.9098456.

A Appendix

Logit Regression Results

Dep. Variable:	KR_LABEL	**No. Observations:**	6695
Model:	Logit	**Df Residuals:**	6686
Method:	MLE	**Df Model:**	8
Date:	Sat, 04 Dec 2021	**Pseudo R-squ.:**	0.2339
Time:	17:22:28	**Log-Likelihood:**	-1175.2
converged:	True	**LL-Null:**	-1534.1
Covariance Type:	nonrobust	**LLR p-value:**	1.056e-149

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-19.8598	3.183	-6.240	0.000	-26.098	-13.622
C(Sex)[T.2]	0.3808	0.121	3.143	0.002	0.143	0.618
KLGrade	1.1399	0.057	19.881	0.000	1.028	1.252
BMI	0.0282	0.012	2.392	0.017	0.005	0.051
Age	0.4353	0.102	4.274	0.000	0.236	0.635
KneeInjury	0.0453	0.123	0.369	0.712	-0.196	0.286
KneeAlignment	-0.0160	0.014	-1.129	0.259	-0.044	0.012
RaceIsBlack	-0.9869	0.174	-5.685	0.000	-1.327	-0.647
I(Age ** 2)	-0.0033	0.001	-4.104	0.000	-0.005	-0.002

Figure 4: Regression output for Baseline Model (1)

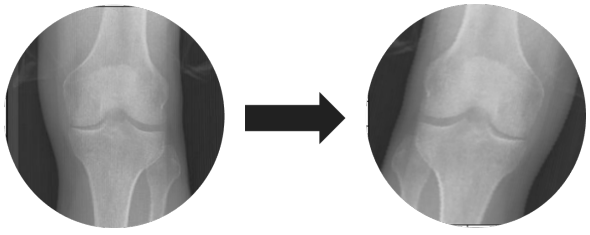

Figure 5: Example X-ray image pre and post data augmentations

Table 6: Feature importances and SHAP values of all features in final CatBoost Model

Feature	LossFuncChgImportance	ShapImportance
KLGrade	0.0137	0.3311
BUkXRReading	0.0035	0.0863
Osteophytes&JSN	0.0033	0.1886
CompositeOAGrade	0.0023	0.1221
FlexionPain	0.0022	0.0789
RaceIsBlack	0.0020	0.1012
AwarenessKneeProblems	0.0019	0.2831
UsedPainMedsL12M	0.0017	0.1829
Age	0.0017	0.3858
KneeCatchHangUpL7D	0.0015	0.0678
DoctorDiagnosedOA	0.0014	0.0889
OtherKnee_KLGrade	0.0014	0.1003
BroughtPrescriptionMedsL30D	0.0013	0.1142
CalciumTumsIntakeL30D	0.0013	0.0873
UsedPainMedEitherKnee	0.0010	0.0628
PhysicalActivityScaleElderly	0.0009	0.0848
ConsideringKneeReplacement	0.0009	0.0224
RxUseRofecoxib	0.0009	0.0225
ChickenTurkeyIntakeL12M	0.0008	0.0822
UsedNSAIDSEveryOtherDayL30D	0.0008	0.0379
BUkXRReadingLateral	0.0008	0.0479
FlexionContractureHyperextension	0.0008	0.0593
DailyCalciumIntake	0.0008	0.0526
KneeDifficultySelfRating	0.0008	0.0602
KneeCapTenderness	0.0007	0.0297
OtherKnee_BUkXRReadingMedial	0.0007	0.0509
UsedPainMedsEveryOtherDayL12M	0.0006	0.0741
OtherKnee_FlexionPain	0.0006	0.0265
TookPainMedsToday	0.0006	0.0184
AbdominalCircumference	0.0005	0.0799
HistoryKneeSurgery	0.0005	0.0355
OtherKnee_CompositeOAGrade	0.0005	0.0503
HoursSitting	0.0004	0.0437
RadialPulse	0.0004	0.0478
FeltFearfulFrequencyL7D	0.0004	0.0564
OtherKnee_MedialJointSpaceNarrowing	0.0003	0.0129
IceCreamConsumedEachTime	0.0003	0.0738
DoctorDiagnosedOtherArthritis	0.0003	0.0629
ChooseLowFatIceCream	0.0003	0.0356
CerealIntakeL12M	0.0002	0.0701
GeneralHealthSelfRated	0.0002	0.0410
OtherKnee_DifficultyStanding	0.0002	0.0371
VegetableIntakeL12M	0.0001	0.0270
DaidzeinIntake	0.0000	0.0209